research articles Journal of proteome research

Amino Acid Sequence Determination of Protein Biomarkers of *Campylobacter upsaliensis* and *C. helveticus* by "Composite" Sequence Proteomic Analysis

Clifton K. Fagerquist*

Western Regional Research Center, Agricultural Research Service, United States Department of Agriculture, 800 Buchanan Street, Albany, California 94710

Received December 28, 2006

We have identified the protein biomarkers observed in the matrix-assisted laser desorption/ionization time-of-flight mass spectra (MALDI-TOF−MS) of cell lysates of five strains of *Campylobacter upsaliensis* and one strain of *C. helveticus* by "bottom-up" proteomic techniques. Only one *C. upsaliensis* strain had previously been genomically sequenced. The significant findings are as follows: (1) The protein biomarkers identified were: 10 kD chaperonin, protein of unknown function (DUF465), phnA protein, probable periplasmic protein, D-methionine-binding lipoprotein MetQ, cytochrome c family protein, DNA-binding protein HU, thioredoxin, asparigenase family protein, helix-turn-helix domain protein, as well as several ribosomal and conserved hypothetical proteins. (2) Amino acid substitutions in protein biomarkers across species and strains account for variations in biomarker ion mass-to-charge (*m/z*). (3) The most common post-translational modifications (PTMs) identified were cleavage of N-terminal methionine and N-terminal signal peptides. The rule that predicts N-terminal methionine cleavage, based on the penultimate residue, does not appear to apply to *C. upsaliensis* proteins when the penultimate residue is threonine. (4) It was discovered that some protein biomarker genes of the genomically sequenced *C. upsaliensis* strain were found to have nucleotide sequences with GTG or TTG "start" codons that were not the actual start codon (ATG) of the protein based on proteomic analysis. (5) Proteomic identification of the protein biomarkers of the non-genomically sequenced *C. upsaliensis* and *C. helveticus* strains involved identification of homologous protein amino acid sequences to that of the sequenced strain. Interestingly, some protein sequence regions that were not completely homologous to the sequenced strain, due to amino acid substitutions, were found to have homologous sequence regions from more phyogenetically distant species/strains, e.g., *C. jejuni*. Exploiting this partial homology of more distant species/strains, it was possible to construct a "composite" amino acid sequence using multiple non-overlapping sequence regions from both phylogenetically proximate and distant strains. The new composite sequence was confirmed by both MS and MS/MS data. Thus, it was possible in some cases to determine the amino acid sequence of an unknown protein biomarker from a genomically non-sequenced bacterial strain without the necessity of either genetically sequencing the biomarker gene or resorting to *de novo* MS/MS analysis of the full protein sequence.

Keywords: *Campylobacter upsaliensis* • *helveticus* • MALDI-TOF-MS • composite sequence • proteomics • post-translational modification • bacterial classification • foodborne pathogen • protein biomarkers

Introduction

Emerging technologies to detect and identify foodborne microorganisms is an active area of analytical science with impact on food safety and public health. A number of analytical approaches are possible, however mass spectrometry-based techniques have gained favor due to their speed, sensitivity, and specificity. The ionization techniques most commonly used in mass spectrometry for protein analysis are electrospray ionization (ESI)[1] and matrix-assisted laser desorption/ionization (MALDI).[2−3] Given the unique chemical and physical processes involved in ionization by ESI and MALDI, selection of the ionization technique is often dictated by the complexity and purity of the type of sample to be analyzed. The ionization technique selected will often determine the type of mass analyzer to be used and the speed of sample analysis. MALDI is traditionally associated with high throughput analysis whereas ESI has been primarily associated with analysis coupled to liquid chromatography (LC), which increases the time of analysis.

Williams, Musser, and co-workers have demonstrated the feasibility of generating a unique protein molecular weight

* To whom correspondence should be addressed. C. K. Fagerquist, Western Regional Research Center, Agricultural Research Service, U.S. Department of Agriculture, 800 Buchanan Street, Albany, CA 94710, USA. E-mail: cfagerquist@pw.usda.gov.

10.1021/pr060700o CCC: $37.00

Published on Web 05/18/2007

(MW) "fingerprint" or profile for bacterial micro-organisms using LC/ESI−MS.[4−6] As each protein generates its own charge state envelope due to the multiple charging nature of ESI, the LC/ESI−MS data must be deconvoluted in time intervals corresponding to the elution of specific proteins. The deconvoluted data is then compiled to provide a unique profile of protein MWs (150+).[4−5] Differences in protein MW profiles among closely related strains (pathogenic versus nonpathogenic) may lead to identification of proteins involved (or at least correlated) to virulence factors. The only significant disadvantage of this approach is the time required for chromatographic separation. To maximize the number of proteins detected (and thus increase the probability of generating a unique protein profile), it is advantageous to maximize chromatographic resolution in order to reduce the possibility of competitive ionization of coeluting proteins. Another challenge is that the dynamic range of protein concentrations in bacteria is several orders-of-magnitude. Significant, but low abundant, proteins may not be detected due to coelution with more abundant proteins. In consequence, good chromatographic resolution is critical but may lengthen the time of analysis.

Another approach that is gaining in popularity for microorganism identification is MALDI time-of-flight mass spectrometry (TOF−MS).[7−19] This technique involves analysis of intact microorganisms (or cell lyates) by detection of high copy primarily cytosolic proteins. As MALDI generates primarily singly charged protein ions, spectral complexity is not a significant problem and no data deconvolution is required. However, the number proteins detected is typically in the range of 20−50, and because the number of open reading frames in a microorganism genome is typically 1500−2500, the number of proteins detected by MALDI-TOF−MS is relatively small. However, even with the relatively small number of proteins detected, MALDI-TOF−MS has demonstrated its utility to differentiate bacterial microorganisms by genus, species, subspecies, and strain.[7−22] In addition, the advantages of this approach are minimal sample preparation (beyond microorganism culturing) and very rapid MS analysis time as no chromatography is involved. MALDI-TOF−MS data are most commonly analyzed by pattern recognition algorithms,[23−24] although increasingly bioinformatics approaches to data analysis are also gaining favor.[25−28]

Our group has extensively analyzed foodborne bacterial microorganisms using MALDI-TOF−MS and, in particular, *Campylobacter*[20−22] which is responsible for 2.4 million incidents of foodborne illness each year in the US.[29] We have demonstrated the ability to speciate[20,21] and even sub-speciate[22] *Campylobacter* using this technique. In addition, we have systematically extracted and identified by "bottom-up" proteomics techniques the proteins responsible for generating the relatively unique protein profile. In the process, we have identified the factors that contribute to making a MS spectrum unique to a particular microorganism.[21,22]

In the current study, we have identified *C. upsaliensis* and *C. helveticus* protein biomarkers that are prominently observed in the MALDI-TOF−MS spectra of bacterial cell lysates. Protein biomarkers of five strains of *C. upsaliensis* were selected for extraction and identification: RM3195, RM2092, RM4249, RM3776, and RM4245. In addition, one strain of *C. helveticus* (a closely related species to *C. upsaliensis*) was also selected for extraction and identification: RM3807. Only the genome of *C. upsaliensis* strain RM3195 has been fully sequenced and annotated,[30] however the other strains in this study have been speciated by extended multi-locus sequence typing.[31] We also demonstrate the ability to determine the full amino acid sequence of a protein biomarker from a genomically non-sequenced bacterial strain by combining proteomic identifications from multiple non-overlapping identifications. This "composite" sequence is then confirmed by MS and MS/MS data.

Materials and Methods

A detailed description of the protocol for protein biomarker detection, extraction, and identification was reported previously.[20−22] The following is an abbreviated description of the experimental procedures employed. *Warning*: *Campylobacter* is a Biosafety Level 2 human pathogen. All appropriate precautions were taken when handling this pathogen. *Campylobacter upsaliensis* and *C. helveticus* strains were grown as previously described[20−22] on nonselective growth media at 37−42 °C for 24−48 h after which cells were harvested and analyzed by MALDI-TOF−MS analysis or prepared for protein extraction and analysis. *Campylobacter* cells were lysed and their proteins extracted using a solution of 2:1 water/acetonitrile, 0.1% trifluoroacetic acid. The cells were lysed by bead-beating in the extraction solution for 60 s. Centrifugation pelleted cellular membrane and other insoluble material. The supernatant was analyzed by MALDI-TOF−MS analysis and separated by high performance liquid chromatography (HP Series II 1090, Palo Alto, CA). HPLC fractions were analyzed by MALDI-TOF−MS. Protein biomarkers in HPLC fractions that were also detected in the original MALDI-TOF−MS analysis of cell lysate were, after centrifugal evaporation, separated by 1-D SDS-PAGE. Gels were fluorescently stained, imaged, and prominent gel bands excised. The gel bands were subjected to in-gel tryptic digestion with an automated protein digester.

MALDI-TOF−MS of cell lysates and HPLC fractions has been described in detail elsewhere.[21,22] Briefly, a sub-saturated solution of *trans*-4-hydroxy-3-methoxy-cinnamic acid (ferulic acid) was prepared in 2:1 water/acetonitrile with 0.1% TFA. The matrix was spotted onto a square stainless-steel target plate with 7 × 7 array of deposition sites. After drying at room temperature, the cell lysate supernatant (or the HPLC eluent) was spotted on top of the dried matrix spot. After external calibration of the instrument, the sample was analyzed on a Reflex II MALDI-TOF mass spectrometer (Bruker Daltonics, Billerica, MS) in reflectron-mode at an ion acceleration voltage of 20 kV resulting in a mass accuracy of ±10 Da. Instrument resolution was 800−1200 fwhm. Data was processed with the software provided with the instrument.

A more accurate molecular weight (MW) measurement was made of a protein biomarker by high-resolution electrospray ionization mass spectrometry (HR-ESI-MS) of the HPLC fraction containing the protein biomarker.[21−22] A quadrupole/time-of-flight (Qq-reflecton-TOF) mass spectrometer (Q-STAR Pulsar I, MDS Sciex/ABI, Toronto, Canada) was utilized for such measurements. The measured MW of a protein biomarker was determined by deconvolution of the charge state envelope of the HR-ESI−MS spectrum using a Bayesian protein reconstruct algorithm available with the commercial software of the instrument (Analyst 1.2). Three separate HR-ESI−MS measurements were made of each biomarker. The reflectron-TOF of the Qq-TOF was calibrated externally with the peptide glufibrogen prior to each measurement.[21−22] An average and standard deviation was then calculated from these three measurements. The TOF resolution was typically FWHM 7000−9000. A discrepancy greater than ±1 Da between the predicted MW and

Table 1. Sources and Locations of Five Strains of *C. upsaliensis* and One Strain of *C. helveticus*

strain	synonym	species	source	location
RM3195	ATCC BAA-1059	*C. upsaliensis*	human	South Africa
RM2092		"	human	unknown
RM4249		"	human	Belgium
RM3776		"	human	South Africa
RM4245		"	human	Belgium
RM3807		*C. helveticus*	feline	USA (CA)

HR-ESI−MS MW suggested the presence of an unidentified PTM or amino acid substitution.

Products of in-gel digestion were analyzed by nanoflow HPLC system (LC Packings/Dionex , Sunnyvale, CA) with a 1000-to-1 split interfaced to the hybrid Qq-TOF instrument. Samples were preconcentrated with a C18 "trap" column prior to being chromatographically separated on the C18 monomeric analytical column. Tryptic peptides of digested proteins were ionized by ESI, detected by a MS survey scan, mass selected by the first quadrupole, fragmented by collision-induced dissociation in the second quadrupole (with nitrogen as the target gas) and the fragment ions analyzed by the TOF mass analyzer. Data were acquired using the data dependent scanning of the instrument software.

The MS/MS data files generated by the instrument software were processed to MGF files using a WIFF-to-DTA batch converter.[32] The DTA files containing MS/MS data were then used to search against a flat file containing amino acid sequences of all eubacterial proteins encoded in the National Center for Biotechnology Information nonredundant database (NCBI nr). An in-house version of MASCOT[33] and Global Proteome Machine[34] search engines were used for database searching. *De novo* sequencing was accomplished using PEAKS software (Bioinformatics Solutions Inc., Version 4.0). The predicted average MW of a protein was calculated from its amino acid sequence using GPMAW software (Lighthouse data, Version 7.0).

Results

***C. upsaliensis* and *C. helveticus* Strains**. Table 1 shows the sources and locations of the five strains of *C. upsaliensis* (RM3195, RM2092, RM4249, RM3776, RM4245) and one strain of *C. helveticus* (RM3807) analyzed in this study. Only *C. upsaliensis* strain RM3195 has been fully genomically sequenced.[30] *C. upsaliensis* strains RM2092, RM4249, RM3776, and RM4245 and *C. helveticus* strain RM3807 were previously characterized by extended multilocus sequence typing (MLST).[31]

MALDI-TOF−MS of *Campylobacter*. Figure 1 shows a typical MALDI-TOF−MS spectrum of a cell lysate extract of *Campylobacter*, specifically of *C. helveticus* strain RM3807. Numerous protein biomarker m/z ion peaks are observed. MALDI-TOF−MS of cell lysates of *C. upsaliensis* strains RM3195, RM2092, RM4249, RM3776, and RM4245 have many more ion peaks (m/z) that are common to each other than are common to the *C. helveticus* strain RM3807 (Tables 2 and 3). Amino acid substitutions caused by non-synonymous mutations of the protein biomarker gene are responsible for "shifts" in ion peaks (m/z) across species and strains.[21−22] The phylogenetic distance between species and strains is reflected in the *number* of biomarkers that undergo a mass shift.

Previously, the protein biomarker ion at m/z 10191 (m/z 5095, +2) in Figure 1, was detected, extracted and proteomically identified to be the DNA-binding protein HU.[21] The HU protein biomarker for strains *C. upsaliensis* strains RM3195, RM2092, RM3776 and RM4245 were also detected, extracted, and identified by proteomic techniques.[21] Proteomic identification relied upon DNA sequencing of the HU gene (*hup*) of a particular strain. In the case of *C. helveticus* strain RM3807, it was not possible to sequence the *hup* gene because the primers used for *C. upsaliensis* did not amplify for the *C. helveticus* strain presumably because the flanking genes were different. However, there was enough conserved homology between *C. upsaliensis* and *C. helveticus* HU such that it was possible to identify the protein biomarker at m/z 10191 as being HU without determining its full sequence. Variations in HU MW among these various strains were the result of the previously mentioned amino acid substitutions caused by nonsynonymous mutations of *hup*.

Protein Biomarker Extraction and Identification. Detection, extraction, and proteomic identification was performed on the other protein biomarkers observed in the MALDI-TOF−MS spectra of RM3195, RM2092, RM4249, RM3776, RM4245, and RM3807. Table 2 and 3 summarizes the protein biomarkers that were identified definitively by proteomic techniques as well as those that could only be assigned tentatively based on limited MS or MS/MS data. Tentative assignments were assigned based on the proximity of the m/z of the biomarker ion in one strain to the m/z of a biomarker ion in another strain that had been identified successfully by proteomics techniques. Mass shifts of ±15−60 Da between biomarkers from different strains/species were assumed to be the result of amino acid substitutions. Including HU, a total of 17 protein biomarkers were identified (although they were not always detected and identified in all strains). They are: DNA-binding protein HU, 50S ribosomal L7/L12, 50S ribosomal L24, protein of unknown function DUF465, asparaginase family protein, 30S ribosomal S16, cytochrome c family protein, hypothetical protein Cup 0937, probable periplasmic protein, 50S ribosomal L29, 10 kD chaperonin, thioredoxin, phnA protein, D-methionine-binding lipoprotein MetQ, 50S ribosomal L27, conserved hypothetical protein, and helix-turn-helix domain protein. Over half of the biomarkers were post-translationally modified. Two of the biomarkers have genes with "start" codons in the nucleotide sequence of strain RM3195 that were incorrect based on proteomic analysis. Proteomic analysis was performed with both MASCOT and GPM search engines (both scores are provided in Tables 2 and 3).

As mentioned previously, only strain RM3195 has been fully genomically sequenced. In consequence, proteomic identification of the protein biomarkers of the other strains in this study were identified based on their full or partial sequence homology to the protein biomarker sequences of RM3195 (or other genomically sequenced *Campylobacter*). In addition to biomarker identification by analysis of MS/MS of tryptic peptides, it was also possible, in some cases, to identify (or confirm) the full amino acid sequence of a protein biomarker by combining multiple identifications (not only the highest scoring identification) to generate a "composite" sequence of the biomarker. A HR-ESI−MS measurement of the biomarker MW was used to confirm the correctness of the composite sequence. The composite sequence was also confirmed by MS/MS data.

Composite Sequence Proteomic Analysis (CSPA) of Bacterial Protein Biomarkers. Table 4 shows the process of identification used to determine the amino acid sequence of a protein biomarker (phnA protein) of *C. upsaliensis* strain RM4245 whose gene is not sequenced. MASCOT and GPM

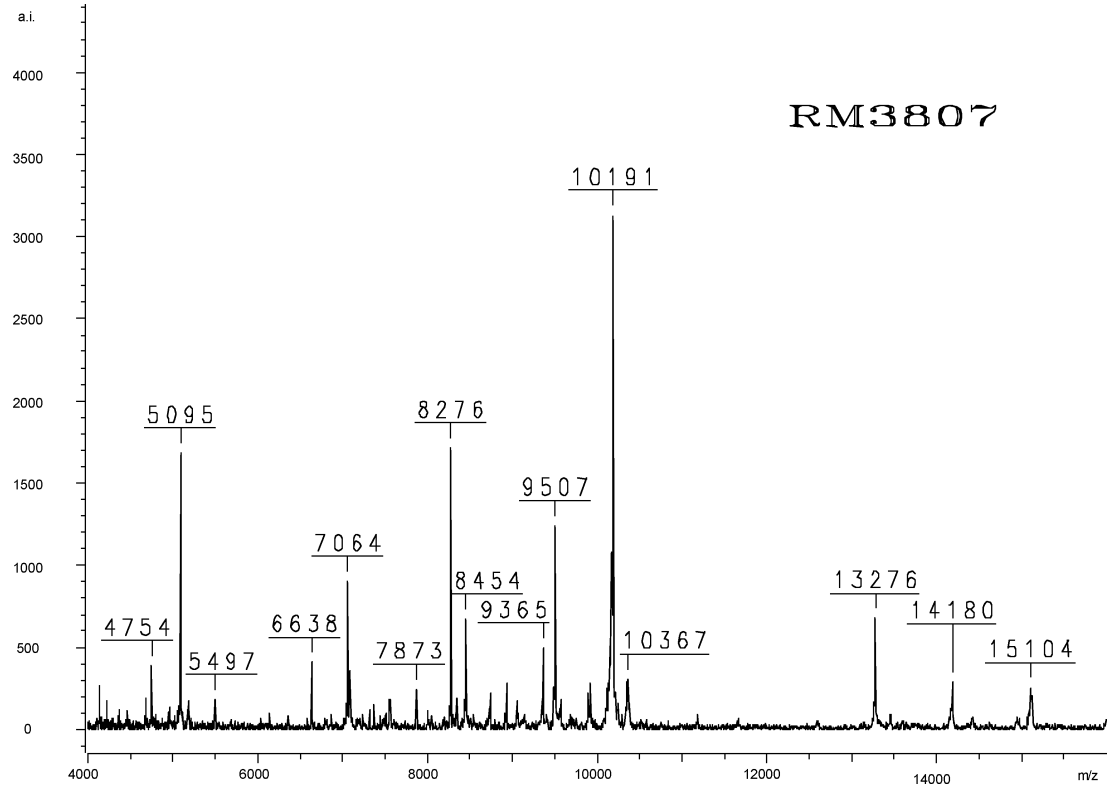

Figure 1. MALDI-TOF−MS of cell lysate of *C. helveticus* strain RM3807.

analysis of its MS/MS datafile return both a high scoring identification (phnA protein of *C. upsaliensis* strain RM3195) and a lower scoring identification (phnA-like protein of *C. jejuni jejuni* NCTC 11168). As RM4245 is an *C. upsaliensis* it is not surprising that the highest scoring identification is another *C. upsaliensis* strain. Interestingly, the lower scoring identification contains sequence coverage of a region of the protein that is not covered by higher scoring identification. Combining the sequence coverage of both the highest and lower scoring identifications results in a composite sequence. The composite sequence also includes protein sequence regions not confirmed by either identification but which are (in this example) common

Table 2. Protein Biomarkers of *C. upsaliensis* Strains RM3195, RM2092, and RM4249 That Were Detected, Extracted, and/or Identified by Proteomic Techniques[a]

		Protein Biomarkers (* signifies presence of a PTM. ** signifies an error in the "start" codon of the biomarker gene.)																
		HU	50S* L7/L12	50S* L24	DUF 465**	Aspara-ginase*	30S S16	Cyto-c*	HP Cup 0937*	Peri-plasmic*	50S L29	Chaper-onin**	Thiore-doxin*	phnA*	D-met MetQ*	50S* L27	CHP	Helix Turn-Helix
RM 3195	MALDI-TOF (m/z)[20]	10272	12856	*7924*	8450	*14207*	8712	9504		13647	7063	9301	11139		13202 (+2)		10022	*9771*
	HR-ESI-MS Ave MW (Da)	10270.8[21] ± 0.3	12855.8 ± 0.2		8449.5 ± 0.2		8712.0 ± 0.1	9501.5 ± 0.0		13648.1 ± 0.3	7061.7 ± 0.1	9299.8 ± 0.1	11137.8 ± 0.2	7350.1 ± 0.3	26413.1 ± 0.7		10026.9 ± 0.3	
	Predicted Ave MW (Da)	10270.9[21]	12986.8 12855.6*	8053.8 (−SH) 7922.6	*11664.5 8449.6**	36333.9 (−SH) 14205.1*	8712.2	10869.8 (3 −SH) 9504.8*	13841.2	15463.7 13647.4*	7061.6	10656.6 9299.8**	11268.0 (S−S, −SH) 11137.8*	7481.7 7350.5*	28568.0 26411.3*	9284.7 9153.7*	10026.5	9768.7
	GPM log(e) Mascot	−155.9[21]	−68.3 884		−26.1 167		−19.6 158	−16.1 225		−70.6 496	−26.1 452	−72.1 553	−43.7 537	−6.2 101	−81.8 951			
RM 2092	MALDI-TOF (m/z)[20]	10171	12886	*7923*	*8422*	*14190*	*8725*	*9503*	13953		*7061*	9317	11137				*10042*	*9787*
	HR-ESI-MS Ave MW (Da)	10171.7[21] ± 0.1	12887.7 ± 0.1						13955.3 ± 0.1									
	Predicted Ave MW (Da)	10171.8[21]		8053.8 (−SH) 7922.6*				10869.8 (3 −SH) 9504.8*			7061.6	9317.9 (Comp.)	11268.0 (S−S, −SH) 11137.8*					
	GPM log(e) Mascot	−91.5[21]	−97.0 957						−32.8 278			−44.6 (Comp.)	−34.2 399					
RM 4249	MALDI-TOF (m/z)	10175	12892	7926	8425	14211	*8729*	*9506*			7064	9321	11170	*7353*	*13185 (+2)*		*10004 10034*	*9797*
	HR-ESI-MS Ave MW (Da)	10171.5 ± 0.2	12887.6 ± 0.3	7922.6 ± 0.1	8421.4 ± 0.1	14204.6 ± 0.2					7061.2 ± 0.2	9318.7 ± 0.7	11165.8 ± 0.2					
	Predicted Ave MW (Da)	10171.8		8053.8 (−SH) 7922.6*		36333.9 (−SH) 14205.1*		10869.8 (3 −SH) 9504.8*			7061.6	9317.9 (Comp.)	11165.8					
	GPM log(e) Mascot	−101.0 472	−43.1 565	−17.7 305	−13.4 119	−87.1 717					−1.3 44	-11.1 (Comp.)	−54.2					

[a] Biomarkers detected and/or extracted from a strain but not identified by proteomics techniques were nonetheless assigned tentatively if the biomarker was identified by proteomics techniques in another strain. Tentative assignments are in italics; the oxidation state of protein cysteines are indicated by S−S or −SH (or both) and incorporated into the MW calculation. Biomarker amino acid sequences and MWs determined by composite sequence analysis and/or *de novo* sequencing are indicated.

Table 3. Protein Biomarkers of *C. upsaliensis* Strains RM3776, RM4245 and *C. helveticus* Strain RM3807 That Were Detected, Extracted, and/or Identified by Proteomic Techniques

		Protein Biomarkers (* signifies presence of a PTM. ** signifies an error in the "start" codon of the biomarker gene.)																
		HU	50S* L7/L12	50S* L24	DUF 465**	Aspara-ginase*	30S S16	Cyto-c*	HP Cup 0937	Peri-plasmic*	50S L29	Chaper-onin**	Thiore doxin*	phnA*	D-met MetQ*	50S* L27	CHP	Helix Turn-Helix
RM 3776	MALDI-TOF (m/z)[20]	10197	12852		8449	*14200*	8712	9502		*13646*	7060	9298	11135		*13205 (+2)*			
	HR-ESI-MS Ave MW (Da)	10198.7 [21] ± 0.1	12855.8 ± 0.3		8449.6 ± 0.2		8712.0 ± 0.2	9501.7 ± 0.1			7061.3 ± 0.1	9299.8 ± 0.4	11137.9 ± 0.3					
	Predicted Ave MW (Da)	10198.9 [21]	12986.8 12855.6*		11664.5 8449.6**	36333.9 (−SH) 14205.1*	8712.2	10869.8 (3 −SH) 9504.8*		15463.7 13647.4*	7061.6	10656.6 9299.8**	11268.0 (S−S,−SH) 11137.8*		28568.0 26411.3*			
	GPM log(e) Mascot	−257.4 [21]	−99.8 1241		−56.3 389		−15.1 440	−29.8 318			−14.3 349	−85.0 618	−40.7 531					
RM 4245	MALDI-TOF (m/z)[20]	10229	12857	*7924*	*8451*	*14209*	*8759*	*9506*		*13663*	*7063*	9302	11139	7325	*13185 (+2)*	9168	10053	9770
	HR-ESI-MS Ave MW (Da)	10226.8 [21] ± 0.2	12855.8 ± 0.1									9299.6 ± 0.3	11137.6 ± 0.3	7323.6 ± 0.1		9181.3 ± 0.0 (Met Ox)	10051.1 ± 0.3	
	Predicted Ave MW (Da)	10226.9 [21]	12986.8 12855.6*	8053.8 (−SH) 7922.6*	11664.5 8449.6**	36333.9 (−SH) 14205.1*		10869.8 (3 −SH) 9504.8			7061.6	10656.6 9299.8**	11268.0 (S−S,−SH) 11137.8*	7454.7 7323.5* (Comp.)				9768.7
	GPM log(e) Mascot	−264.8 [21]	−112.5 1225									−88.3 678	−84.7 680	−30.0 (Comp.)		−64.2 398	−54.2 590	154
RM 3807	MALDI-TOF (m/z)	10191 [21]	13276		*8454*	14180	*8741*	*9507*			7064	9365	11179	*7376*				
	HR-ESI-MS Ave MW (Da)	10186.9 [21] ± 0.3	13270.9 ± 0.4			14175.1 ± 0.4					7061.3 ± 0.2	9361.3 ± 0.6	11177.0 ± 0.4					
	Predicted Ave MW (Da)	10185.9 (Comp.& *de novo*)			11664.5 8449.6**			10869.8 (3 -SH) 9504.8*			7061.6	9361.9 (Comp.& *de novo*)	11176.8 (Comp. & *de novo*)					
	GPM log(e) Mascot	−186.4 (Comp.& *de novo*)	−139.7 989			−26.0 127					−41.6 469	−61.0 (Comp.& *de novo*)	−68.3 (Comp.& *de novo*)					

[a] Biomarkers detected and/or extracted from a strain but not identified by proteomics techniques were nonetheless assigned tentatively if the biomarker was identified by proteomics techniques in another strain. Tentative assignments are in italics. The oxidation state of protein cysteines are indicated by S−S or −SH (or both) and incorporated into the MW calculation. Biomarker amino acid sequences and MWs determined by composite sequence analysis and/or *de novo* sequencing are indicated.

to both sequences. The MW of this composite sequence is then calculated and compared to the HR-ESI−MS MW of the protein biomarker. Excellent agreement was obtained between the predicted MW and HR-ESI−MS MW when N-terminal methionine cleavage is included. As the penultimate residue of this protein is alanine, post-translational removal of methionine is consistent with the bacterial rule that predicts N-terminal methione cleavage.[26,35−37] Finally, the composite sequence is included into the GPM database and the MS/MS datafile is reanalyzed. The result is an improved identification score and greater sequence coverage of the protein biomarker. MS and MS/MS confirmation of the composite sequence leaves little doubt as to its correctness. Thus, it is possible, in some cases, to identify the full amino acid sequence of a protein biomarker of a genomically nonsequenced bacterial strain using CSPA. A brief comment about terminology. The term "consensus sequence" is commonly used in genomics and bioinformatics to refer to the process by which a number of different sequences (usually DNA sequences) are aligned, and the common elements are highlighted or selected. In contrast, the process of constructing a composite sequence is not simply combining "common" elements from differing amino acid sequences (although this may also occur). A composite sequence specifically includes non-overlapping sequence regions that are *not* common to all sequences but which are determined to be correct based on MS/MS data and database searching. Thus, a composite sequence is not the same as a "consensus sequence" as the processes involved are actually different.

A more complex example using CSPA is shown in Table 5 for the identification of 10 kD chaperonin of *C. helveticus* strain RM3807. MASCOT and GPM analysis of its MS/MS datafile returned both a high scoring identification (10 kD chaperonin of *C. upsaliensis* strain RM3195) and a lower scoring identification (10 kD chaperonin of *C. jejuni jejuni* strain RM1221). Interestingly, the high scoring identification also revealed a problem with amino acid sequence of 10 kD chaperonin of RM3195 in the NCBI database. Specifically, the N-terminus of the protein is not confirmed by proteomic analysis. In addition, the "start" codon of the chaperonin gene (*groES*) for RM3195 is "GTG" a possible (although not typical) start codon of a bacterial gene sequence. In any case, combining the high and low scoring identifications results in a composite sequence covering over 75% of the biomarker sequence. However, there still remain sequence regions unconfirmed by MS/MS that are common (and not common) to both identifications. Calculating the predicted MW using either of these possibilities results in a value that is not consistent with the HR-ESI−MS MW of the biomarker. *De novo* analysis of the MS/MS data and focusing primarily on identifying those sequence regions not confirmed by database identifications resulted in identification of the sequence ^{46}EVSDVTSGDKILFAK (and another sequence ^{10}-VLVK). The ^{46}EVS... *de novo* sequence although more similar to the *C. upsaliensis* sequence than to *C. jejuni* sequence is still different by three residues (boxed). Addition of this partial sequence to the composite sequence results in a full sequence with a predicted MW that is in good agreement with the HR-ESI−MS MW of the biomarker. Finally, confirmation of this combined composite and *de novo* sequence by GPM resulted in a significantly higher score and almost complete confirmation of the entire biomarker sequence. Protein biomarker sequences determined by composite sequencing or composite and *de novo* sequencing are indicated in Tables 2 and 3.

Post-Translational Modifications of *C. upsaliensis* and *C. helveticus* Protein Biomarkers. The post-translational modification (PTM) most commonly detected and identified was N-terminal methionine (Nt-Met) cleavage. The presence of this PTM was confirmed by MS and/or MS/MS for the following proteins: thioredoxin, phnA protein, and the 50S ribosomal proteins L7/L12, L24, and L27. The penultimate residue of these proteins is either alanine or glycine. The rule that governs Nt-

Table 4. Proteomic Analysis of the phnA Protein from *C. upsaliensis* Strain RM4245[a]

MASCOT Analysis of RM4245 hplc#48 gs1 060616

Score: 276

phnA protein [*Campylobacter upsaliensis* RM3195]

1 MAKDSNGTELSAGDSVSVIKDLKVKGANTTLKRGTTIKNIKLTNKDSEIE
51 AKVDKFGTLVLKTEFLKKI

Score: 60

phnA-like protein [*Campylobacter jejuni* subsp. *jejuni* NCTC 11168]

1 MAKDANGIELNAGDSVSIIKDLKVKGASTTLKRGTTIKNIKLTSKEGEIE
51 ARVDKFGVIVLKTEFLKKI

GPM Analysis of RM4245 hplc#48 gs1 060616

Log(e) = -22.0

Cu1607 (*Campylobacter upsaliensis* RM3195)

1 MAKDSNGTELSAGDSVSVIKDLKVKGANTTLKRGTTIKNIKLTNKDSEIE
51 AKVDKFGTLVLKTEFLKKI

Log(e) = -2.3

NCTC 11168 Cj0185c_[Conserved_hypothetical_protein,_PhnA-like_protein]

1 MAKDANGIELNAGDSVSIIKDLKVKGASTTLKRGTTIKNIKLTSKEGEIE
51 ARVDKFGVIVLKTEFLKKI

Composite Sequence

1 MAKDSNGTELSAGDSVSVIKDLKVKGASTTLKRGTTIKNIKLTNKDSEIE
51 AKVDKFGTLVLKTEFLKKI MW = 7454.7 Da (w/o N-terminal Met MW = 7323.5 Da).
HR-ESI-MS MW = 7323.6 ± 0.1 Da

Confirmation of Composite Sequence by GPM Analysis

log(e) = -30.0

RM4245_phnA_[phnA_protein_composite_RM3195_ NCTC_11168]

1 MAKDSNGTELSAGDSVSVIKDLKVKGASTTLKRGTTIKNIKLTNKDSEIE
51 AKVDKFGTLVLKTEFLKKI

[a] Boxed residues highlight amino acid variations between high and low scoring identifications (*C. upsaliensis* strain RM3195 and *C. jejuni jejuni* strain NCTC 11168, respectively) using both MASCOT and GPM analysis. Red indicates the sequence was confirmed by MS/MS. The composite sequence combines non-overlapping sequence regions of both the highest and lower scoring identifications that were confirmed by MS/MS. The composite sequence also includes unconfirmed sequence regions common to both strains.

Met cleavage for bacterial proteins, as determined by the penultimate residue, predicts Nt-Met removal when the penultimate residue is alanine or glycine.[26,35−37] Thus, rule and observation are consistent. The rule also predicts Nt-Met removal when the penultimate residue is threonine, however it was observed that both the DNA-binding protein HU and the 30S ribosomal S16 protein retain their Nt-Met even though threonine is the penultimate residue. Our previously reported results have also found this to be the case for *C. jejuni*, *C. coli*, *C. lari*, and *C. concisus* strains.[21,22] Previously, we conjectured that the methionine aminopeptidase responsible for Nt-Met removal has a somewhat different activity in *Campylobacter* species compared to other bacteria.[22] Differences in methionine aminopeptidase activity among different bacteria are due presumably to differences in the primary sequence of the enzyme. For example, the primary sequence of methionine aminopeptidase for *E. coli* O157:H7 strain Sakai is significantly different from the primary sequence of *C. upsaliensis* strain RM3195.

Other PTMs were also detected in four protein biomarkers: asparaginase family protein, probable periplasmic protein, cytochrome c family protein, and D-methionine-binding lipo-

Table 5. Proteomic Analysis of the 10 kD Chaperonin from *C. helveticus* Strain RM3807

[a] Boxed residues highlight amino acid variations between highest and lower scoring identifications (*C. upsaliensis* strain RM3195 and *C. jejuni jejuni* strain RM1221, respectively) using both MASCOT and GPM analysis. Red indicates the sequence was confirmed by MS/MS. The composite sequence combines sequence regions of both high and low scoring identifications that were confirmed by MS/MS as well as unconfirmed sequence of both strains. These unconfirmed sequence regions in the composite sequence did not correspond to HR-ESI-MS MW of the biomarker. In consequence, *de novo* sequencing (in blue) was employed to fill in "gaps" in the sequence.

protein MetQ. These proteins were found to have undergone post-translational removal of a signal peptide or cleavage of the polypeptide chain as shown in Table 6. In the case of cytochrome c family protein, in addition to removal of a 19-residue signal peptide from the N-terminus, there is also covalent attachment of a heme group to the polypeptide *via* two thioether bonds with two cysteine residues. This protein biomarker and its PTMs were also detected in *C. jejuni* strains.[22] A ~ 3 Da discrepancy between the measured and predicted protein MW is due presumably to another PTM. A similar mass discrepancy was observed in *C. jejuni* cytochrome c-553 except there the mass discrepancy was ~ 2 Da.[22] In the case of the asparaginase family protein, the protein biomarker detected, extracted, and identified may (or may not) be the functional protein. The protein appears to be severed in half with only the C-terminal polypeptide being detected (but detected consistently) in several strains.

"Start" Codon Discrepancies in the Protein Biomarker Genes of *C. upsaliensis* Strain RM3195. Proteomic analysis of two protein biomarkers: 10 kD chaperonin and domain of unknown function protein (DUF465), revealed that the actual N-terminus of the protein was not consistent with the sequence shown in the NCBI nr database for *C. upsaliensis* strain RM3195. As previously mentioned, the nucleotide "start" codon for 10 kD chaperonin gene (*groES*) was "GTG" (not the commonly expected "ATG"). Similarly, the actual N-terminus of the protein DUF465 is not consistent with the amino acid sequence in the database and its gene (locus tag = "CUP0317")

Table 6. Examples of Four Protein Biomarkers Detected, Extracted, and Identified from *C. upsaliensis* and *C. helveticus* Strains That Were Shown to Have Undergone Post-Translational Removal of a Signal Peptide or Cleavage of the Polypeptide Chain

RM4249 Asparaginase Family Protein

1 MNLVKGGKKALLVMASLAFLSSVNLYAKDFKPVIVIHGGTSGLGLTKEEFAKREMVMKESLKAGQKVLEAGGSSTDAVIA
81 AIKVMEDSPEFNAGKGAVFTADGYNELDASLMDGKTLNAGAIAMARTIKNPIEAAKVVMEKTPHTLIAGEGADKLAKANG
161 LEIVKQKYFYTEHRYKQLKEAQKSKEILLDSDKAKAHLGLSTEPYLGTVGAIALDKNGNLAAGTSTGGTTNKMTGRIGDS
241 PIIGAGNYANNDSVAVSCTGTGDIYIRVAAAHEVAALYKHKKLSIQKAAEETIKEVAELGGTGGIISIDKNGKVGYAWTK
321 DKLGMYHGEARLGEEPKIFWPVAK MW = 36333.9 Da

161 [208]TVGAIALDKNGNLAAGTSTGGTTNKMTGRIGDS
241 PIIGAGNYANNDSVAVSCTGTGDIYIRVAAAHEVAALYKHKKLSIQKAAEETIKEVAELGGTGGIISIDKNGKVGYAWTK
321 DKLGMYHGEARLGEEPKIFWPVAK MW = 14205.1 Da (HR-ESI-MS MW = 14204.6 ± 0.2 Da)

RM3195 D-methionine-binding lipoprotein MetQ

1 MNIKTLLLVSLVGFTLNLNAAEKIVVAATPVPHAEILNQAKEELKKEGYTLEVKEFTDYVLPNLATDNSEVDANFFQHAP
81 YLEEFNKSKGTKLVKVANIHIEPMAVYSKKYKNFNELKDGAKIAVPNDPTNESRALDIIAKTGLVGFNDKVLKTPIDITQ
161 NPKNIKFIELKAAQLPRALDDVDVAVINSNYALSANLNPVKDSIFIEDKDSPYANILVVKEGKEQDPKIKALTKALQSEK
241 IKKFIEEKYNGAVIPVF MW = 28568.0 Da

1 [20]AEKIVVAATPVPHAEILNQAKEELKKEGYTLEVKEFTDYVLPNLATDNSEVDANFFQHAP
81 YLEEFNKSKGTKLVKVANIHIEPMAVYSKKYKNFNELKDGAKIAVPNDPTNESRALDIIAKTGLVGFNDKVLKTPIDITQ
161 NPKNIKFIELKAAQLPRALDDVDVAVINSNYALSANLNPVKDSIFIEDKDSPYANILVVKEGKEQDPKIKALTKALQSEK
241 IKKFIEEKYNGAVIPVF MW = 26411.3 Da (HR-ESI-MS MW = 26413.1 ± 0.7)

RM3195 Cytochrome c family protein

1 MKKFVAISVLACLAISAYAADGATLYKKC*AVC*HGPKADKVYLNKVPALKSISTAERLQYMKEYSEGKRNAYGQGAIMKIN
81 LKGLTEEDFKALEAYIETL MW = 10869.8 Da

1 [20]ADGATLYKKC*AVC*HGPKADKVYLNKVPALKSISTAERLQYMKEYSEGKRNAYGQGAIMKIN
81 LKGLTEEDFKALEAYIETL MW = 8888.3 Da (+ Heme = 9504.8 Da)
HR-ESI-MS MW = 9501.5 ± 0.0 Da

RM3195 Probable periplasmic protein

1 MKTKIALAALLSAGVLYAQDIVISPAQLPNNAQSFLSTHFKGVNIGLAKKDLDSYDVTLTDGTEVDFIINGEWKEVDGKY
81 KAIPLGFLPSTLVAKVQVSQPNAQIFKVEKQINGYKFKFNNNMEVYTDFNGNILGQKFDD MW = 15463.7 Da

1 [19]QDIVISPAQLPNNAQSFLSTHFKGVNIGLAKKDLDSYDVTLTDGTEVDFIINGEWKEVDGKY
81 KAIPLGFLPSTLVAKVQVSQPNAQIFKVEKQINGYKFKFNNNMEVYTDFNGNILGQKFDD MW = 13647.4 Da
HR-ESI-MS MW = 13648.1 ± 0.3 Da

[a] Signal peptide or polypeptide removed is boxed. Sequence confirmed by MS/MS is highlighted in black. Predicted MW of truncated polypeptide is compared to the HR-ESI-MS MW of the biomarker.

also has a "TTG" "start" codon. Table 7 shows the amino acid sequence of selected protein biomarkers analyzed in this study. The NCBI nr amino acid sequence for 10 kD chaperonin and DUF465 are shown (with accession numbers) including that part of the sequence that was confirmed by MS and MS/MS. Table 7 also aligns amino acid sequences of other selected

Table 7. Amino Acid Sequences of Protein Biomarkers of *C. upsaliensis* Strains RM3195, RM2092, RM4249, RM3776, RM4245 and *C. helevticus* RM3807

DNA-binding protein HU (*hup*)

RM3195	MTKADFISQVAQNAELTKKDAGAATDAVISTITEVLAKGDSISFIGFGTFSTTERAAREARVPSTGKTIKVPATRVAKFKVGKNLKEAVAAKAGKKKK	(AAW83383)[21]
RM2092	MTKADFISQVAQNAGLTKKDAGAATDAVISTITEVLAKGDSISFIGFGTFSTTERAAREARVPSTGKTIKVPATRVAKFKVGKSLKEAVAAKAGKKKK	(AAW83364)[21]
RM4249	MTKADFISQVAQNAGLTKKDAGAATDAVISTITEVLAKGDSISFIGFGTFSTTERAAREARVPSTGKTIKVPATRVAKFKVGKSLKEAVAAKAGKKKK	(AAW83463) (Sequence confirmed by MS & MS/MS).
RM3776	MTKADFISQVAQNAGLTKKDAGAATDAVISTITEVLAKGDSISFIGFGTFSTTERAAREARVPSTGKTIKVPATRVAKFKVGKNLKEAVAAKAGKKKK	(AAW83401)[21]
RM4245	MTKADFISQVAQNAGLTKKDAGAATDAVISTITEVLAKGDSISFIGFGTFSTTERAAREARVPSTGKTIKVPATRVVKFKVGKNLKEAVAAKAGKKKK	(AAW83460)[21]
RM3807	MTKADFISLVAQTAGLTKKDATAATDAVISTITEVLAKGDSISFIGFGTFSTTERAARREARVPSTGATIKVPATRVVKFKVGKNLKEAVAAKAGKKKK	(Non-underlined composite/*de novo* sequence confirmed by MS/MS)

10 kD chaperonin (*groES*)**

RM3195	MLKINILRMDKMNFQPLGKRVLVKRVEETKTTASGIIIPDNAKEKPLIGEVVAVSKEVSDIASGDKIVFAKYGGTEVKLNDGEYLVLNLDDVLGILK	(ZP_00370617)
RM3195	MNFQPLGKRVLVKRVEETKTTASGIIIPDNAKEKPLIGEVVAVSKEVSDIASGDKIVFAKYGGTEVKLNDGEYLVLNLDDVLGILK	(Sequence confirmed by MS & MS/MS)
RM2092	MNFQPLGKRVLVKRVEETKTTASGIIIPDNAKEKPLMGEVVAVSKEVSDIASGDKIVFAKYGGTEVKLNDGEYLVLNLDDVLGILK	(Composite sequence confirmed by MS & MS/MS)
RM4249	MNFQPLGKRVLVKRVEETKTTASGIIIPDNAKEKPLMGEVVAVSKEVSDIASGDKIVFAKYGGTEVKLNDGEYLVLNLDDVLGILK	(Composite sequence confirmed by MS & MS/MS)
RM3776	MNFQPLGKRVLVKRVEETKTTASGIIIPDNAKEKPLIGEVVAVSKEVSDIASGDKIVFAKYGGTEVKLNDGEYLVLNLDDVLGILK	(Sequence confirmed by MS & MS/MS)
RM4245	MNFQPLGKRVLVKRVEETKTTASGIIIPDNAKEKPLIGEVVAVSKEVSDIASGDKIVFAKYGGTEVKLNDGEYLVLNLDDVLGILK	(Sequence confirmed by MS & MS/MS)
RM3807	MNFQPLGKRVLVKRVEETKTTASGIIIPDNAKEKPLMGEVVAVSKEVSDVTSGDKILFAKYGGTEIKLNDGEYLVLNLDDVLGILK	(Composite/*de novo* sequence confirmed by MS and MS/MS)

Thioredoxin (*trx*)*

RM3195	MGKYIELTSENFATAKEGVALVDFWAPWCGPCKMLSPVIDELASDFEGKAKICKVNTDEQGDLAAEYGVRSIPTLIFFKNGEVVGQLVGAQSKQTIADKINSLL	(ZP_00372057, confirmed sequence by MS & MS/MS)
RM2092	MGKYIELTSENFATAKEGVALVDFWAPWCGPCKMLSPVIDELASDFEGKAKICKVNTDEQGDLAAEYGVRSIPTLIFFKNGEVVGQLVGAQSKQTIADKINSLL	(Sequence confirmed by MS and MS/MS)
RM4249	MGKYIELTSENFATAKEGVALVDFWAPWCGPCRMLSPVIDELASDFEGKAKICKVNTDEQGDLAAEYGVRSIPTLIFFKNGEVVGQLVGAQSKQTIADKINSLL	(Sequence confirmed by MS & MS/MS)
RM3776	MGKYIELTSENFATAKEGVALVDFWAPWCGPCKMLSPVIDELASDFEGKAKICKVNTDEQGDLAAEYGVRSIPTLIFFKNGEVVGQLVGAQSKQTIADKINSLL	(Sequence confirmed by MS and MS/MS)
RM4245	MGKYIELTSENFATAKEGVALVDFWAPWCGPCKMLSPVIDELASDFEGKAKICKVNTDEQGDLAAEYGVRSIPTLIFFKNGEVVGQLVGAQSKQTIADKINSLL	(Sequence confirmed by MS and MS/MS)
RM3807	MGKYIELTSENFATAKEGVALVDFWAPWCGPCRMLAPVIDELANDFDGKAKICKVNTDEQGDLAAEYGVRSIPTLIFFKNGEVVAQLVGAQSKQTIADKINSLL	(Composite/*de novo* sequence conf. by MS & MS/MS)

Protein of Unknown Function DUF465**

RM3195	MKIFKQKLYFFCYNLAFNIQNKKGVFMLHEFRDLMSELKGKDAHFDKLFERHNELDDKIKDAEEGRAFLSDVEISTLKKEKLHVKDELAQYLANYKK	(ZP_00370213)
RM3195	MLHEFRDLMSELKGKDAHFDKLFERHNELDDKIKDAEEGRAFLSDVEISTLKKEKLHVKDELAQYLANYKK	(Sequence confirmed by MS & MS/MS)
RM2092	(Detected by MALDI-TOF-MS, but identity not confirmed).	
RM4249	MLHEFRDLMSELKGKDAHFDKLFERHNELDDKIKDAEEGRAFLSDVEISTLKKEKLHVKDELAQYLANYKK	(Non-underlined sequence confirmed by MS/MS)
RM3776	MLHEFRDLMSELKGKDAHFDKLFERHNELDDKIKDAEEGRAFLSDVEISTLKKEKLHVKDELAQYLANYKK	(Sequence confirmed by MS & MS/MS)
RM4245	MLHEFRDLMSELKGKDAHFDKLFERHNELDDKIKDAEEGRAFLSDVEISTLKKEKLHVKDELAQYLANYKK	(Detected by MALDI-TOF-MS, but identity not confirmed)
RM3807	MLHEFRDLMSELKGKDAHFDKLFERHNELDDKIKDAEEGRAFLSDVEISTLKKEKLHVKDELAQYLANYKK	(Detected by MALDI-TOF-MS, but identity not confirmed)

phnA protein*

RM3195	MAKDSNGTELSAGDSVSVIKDLKVKGANTTLKRGTTIKNIKLTNKDSEIEAKVDKFGTLVLKTEFLKKI	(ZP_00371593) (Not detected by MALDI-TOF-MS but sequence confirmed by MS & MS/MS)
RM2092	Not detected.	
RM4249	Detected by MALDI-TOF-MS but identity not confirmed.	
RM3776	Not detected.	
RM4245	MAKDSNGTELSAGDSVSVIKDLKVKGASTTLKRGTTIKNIKLTNKDSEIEAKVDKFGTLVLKTEFLKKI	(Composite sequence confirmed by MS & MS/MS)
RM3807	Detected by MALDI-TOF-MS but identity not confirmed.	

[a] Only RM3195 has undergone genomic sequencing. Boxed residues highlight amino acid variations among species/strains. Underlined residues indicate uncertainty in the sequence due to absence of MS/MS confirmation and discrepancies between theoretical protein MW and biomarker HR-ESI-MS MW. The NCBI GenBank accession numbers of the biomarker genes are given in parentheses next to the sequence. One asterisk indicates the protein biomarker has a PTM. Two asterisks indicates that the gene of the protein has a "start" codon that is not confirmed by MS and MS/MS.

protein biomarkers. Amino acid residue variations across species/strains are boxed.

Discussion

Variations in Protein Biomarker MW Across Species/Strains of *C. upsaliensis* and *C. helveticus*. As shown in Table 7, variations in protein biomarker MW observed across species/strains are the result of amino acid substitutions. The greater the phylogenetic distance between species and strains, the greater the number of amino acid variations identified among their protein biomarkers, and the greater number of mass "shifts" detected in MALDI-TOF−MS spectra. PTMs do not appear to contribute to protein biomarker MW variations as they are consistently observed across species and strains, at least, in *Campylobacter*.[21,22] As shown in Table 7, there are many more amino acid subsitutions between the *C. helveticus* strain and the five *C. upsaliensis* strains than among the *C. upsaliensis* strains. Although *C. helveticus* is phylogenetically closer to *C. upsaliensis* than to other *Campylobacter* species, the protein biomarkers of RM3807 are significantly different from the protein biomarkers of the *C. upsaliensis* strains in this study. Only three (of the ten) biomarker MWs of RM3807 are common to most (although not all) of the protein biomarker MWs of the *C. upsaliensis* strains: cytochrome c family protein, 50S ribosomal L29 protein, and DUF465. Thus, it is not surprising that strain RM3807 is from a different (although closely related) species than the other strains analyzed in this study. Similar speciation of *Campylobacter* strains, on the basis of biomarker MWs, was reported by our laboratory for *Campylobacter*,[20−21] as well as sub-speciation of *C. jejuni* such that it was possible to differentiate *C. jejuni* subsp. *jejuni* from *C. jejuni* subsp. *doylei*.[22]

Determining the Amino Acid Sequence of a Bacterial Protein Biomarker by CSPA. In the course of this work, the full amino acid sequence of unknown protein biomarkers from genomically nonsequenced bacterial strains were determined without DNA sequencing of the biomarker gene or full *de novo* MS/MS sequencing. This was accomplished by careful examination of the proteomic identifications, specifically the highest scoring identification as well as lower scoring identifications. Occasionally, a lower scoring identification confirmed a region of the biomarker sequence that was not confirmed by the highest scoring identification. Combining sequence confirmations from both the highest scoring identification and lower scoring identifications resulted, as previously noted, in a composite sequence that could then be confirmed by MS and MS/MS analysis. What was surprising about this process was that lower scoring identifications were often of proteins from different species of *Campylobacter*, e.g. *C. jejuni*. Thus, some of the protein biomarkers of *C. upsaliensis* strains RM2092, RM4249, RM3776, and RM4245 and *C. helveticus* strain RM3807 had sequence regions that were more consistent with a *C. jejuni* strain than with *C. upsaliensis* strain RM3195. Presumably, this is due to inter-species transfer of genetic information or an artifact of bacterial evolution.

The increasing number of bacterial genomic sequences available in public databases make CSPA increasingly feasible. For instance, in the determination of the sequence of the DNA-binding protein HU for *C. helveticus* strain RM3807, a MASCOT search resulted in three identifications from *C. jejuni*, *C. upsaliensis* and *C. lari* strains with scores of 163, 116, and 83, respectively (see Supporting Information). However, the three identifications covered non-overlapping regions of the HU sequence. A composite sequence was constructed from the three identifications. The composite sequence contained a significant gap that was not caused by the presence of basic residues that might generate very low MW tryptic peptides. *De novo* sequencing was then used to fill in the "gap", which was made easier by the fact that the entire protein sequence was not being determined but only the "gap". Although it was possible to use *de novo* sequencing at the outset of the analysis, the task of sequence confirmation was made significantly easier by performing CSPA prior to *de novo* sequencing. In addition, CSPA may be sufficient for complete determination of the sequence without the necessity of *de novo* sequencing. The combined composite/*de novo* sequence was then compared to the HR-ESI−MS MW of the protein biomarker. If concurrence was obtained, then the new composite/*de novo* sequence was added to the database and confirmed by GPM analysis. In the case of RM3807 HU, GPM analysis could not confirm the entire composite/*de novo* sequence and there continued to be a 1 Da discrepancy between the predicted MW and the HR-ESI−MS MW of the biomarker.

Advantages of CSPA. One clear advantage of CSPA is that, in determining the full amino acid sequence of a protein biomarker from a genomically non-sequenced bacterial strain, it facilitates the reverse engineering of the biomarker gene. Once the full amino acid sequence is known, degenerate internal primers can be designed to sequence outward from the gene to obtain the flanking sequences surrounding the biomarker gene. Once these flanking sequences are obtained, it is then possible to design primers to sequence inward to obtain the full biomarker gene sequence. This approach is particularly useful when the biomarker gene is from a nonsequenced emerging bacterial strain/species whose flanking sequences may be significantly different from existing bacterial genomes. Using primers from known genomes may result in no amplification for the emerging strain. Thus, CSPA can determine both the existence of the gene and facilitate a PCR approach to determining the flanking sequences and ultimately the full genetic sequence of the biomarker.

Errors in Start Codon of Protein Biomarker Genes in Genome of *C. upsaliensis* Strain RM3195. Two of the 17 protein biomarker genes of *C. upsaliensis* strain RM3195 had an incorrect start codon. As the protein biomarkers were selected for proteomic identification on the basis of their appearance in cell lysates analyzed by MALDI-TOF−MS, i.e., randomly, it is likely that there are many more start codon errors in this genomic sequence. As there are ~1940 genes in the genome of *C. upsaliensis* strain RM3195, this may suggest that there may be as many as ~200 genes in the RM3195 genome with similar start codon errors. It is interesting to observe how proteomics techniques can be used, not only to identify PTMs, but also to correct errors in the genomic sequencing.

Conclusions

Seventeen protein biomarkers from five strains of *C. upsaliensis* and one strain of *C. helveticus* have been detected, extracted, and identified by proteomic techniques. Variations in biomarker MW across species/strains were the result of amino acid substitutions. PTMs were consistent across species/strains of *Campylobacter* and thus not responsible for biomarker MW variations across species/strains. However, the number of PTMs detected in these 17 proteins suggest that algorithms used to "identify" protein biomarkers exclusively

by comparison of genomically derived protein MW to MALDI-TOF m/z may result in mis-identification of the protein and lowered confidence in the microorganism identification. The rule that predicts N-terminal methionine cleavage for bacterial proteins, as determined by the penultimate residue, was not followed for *C. upsaliensis* and *C. helveticus* strains when the penultimate residue was threonine. Two of the 17 protein biomarker genes of *C. upsaliensis* strain RM3195 were found to have an incorrect start codon as determined by proteomic analysis. CSPA is a valuable technique for determining the full amino acid sequence of a protein biomarker from a bacterial strain whose genome has not been sequenced. As the number of bacterial genomes increase, this technique has the potential to identify the full amino acid sequences of many proteins which would not be possible without full genomic sequencing.

Acknowledgment. We thank Anna H. Bates and Felicidad Bautista for culturing the strains used in this study. We thank Dr. William H. Vensel for useful discussions. Mention of a brand or firm name does not constitute an endorsement by the U.S. Department of Agriculture over other of a similar nature not mentioned. This article is a U.S. Government work and is in the public domain in the U.S.A.

Supporting Information Available: Search engine parameters and protein biomarkers. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. *Science* **1989**, *246* 64−71.
(2) Karas, M.; Bachmann, D.; Bahr, U.; Hillenkamp, F. *Int. J. Mass Spectrom. Ion Processes* **1987**, *78* 53−68.
(3) Tanaka, K.; Ido, Y.; Akita, S.; Yoshida, Y.; Yoshida, T. *Second Japan-China Joint Symposium on Mass Spectrometry* (abstract), Osaka, Japan, September 15−18, 1987.
(4) Williams, T. L.; Monday, S. R.; Edelson-Mammel, S.; Buchanan, R.; Musser, S. M. *Proteomics* **2005**, *5*, 4161−4169.
(5) Williams, T. L.; Monday, S. R.; Feng, P. C. H.; Musser, S. M. *J. Biomol. Tech.* **2005**, *16*, 134−142.
(6) Williams, T. L.; Musser, S. M.; Nordstrom, J. L.; DePaola, A.; Monday, S. R. *J. Clin. Microbiol.* **2004**, *42,* 1657−1665.
(7) Cain, T. C.; Lubman, D. M.; Weber, W. J., Jr. *Rapid Commun. Mass Spectrom.* **1994**, *8*, 1026−1030.
(8) Krishnamurthy, T.; Ross, P. L.; Rajamani, U. *Rapid Commun. Mass Spectrom.* **1996**, *10,* 883−888.
(9) Krishnamurthy, T.; Ross, P. L. *Rapid Commun. Mass Spectrom.* **1996**, *10,* 1992−1996.
(10) Holland, R. D.; Wilkes, J. G.; Rafii, F.; Sutherland, J. B.; Persons, C. C.; Voorhees, K. J.; Lay, J. O., Jr. *Rapid Commun. Mass Spectrom.* **1996**, *10,* 1227−1232.
(11) Arnold, R.; Reilly, J. *Rapid Commun. Mass Spectrom.* **1998**, *12,* 630−636.
(12) Welham, K.; Domin, M.; Scannell, D.; Cohen, E.; Ashton, D. *Rapid Commun. Mass Spectrom.* **1998**, *12*, 176−180.
(13) Haag, A.; Taylor, S.; Johnston, K.; Cole, R. *J. Mass Spectrom.* **1998**, *33*, 750−756.
(14) Wang, Z.; Russon, L.; Li, L.; Roser, D.; Long, S. R. *Rapid Commun. Mass Spectrom.* **1998**, *12*, 456−464.
(15) Dai, Y.; Li, L.; Roser, D.; Long, S. R. *Rapid Commun. Mass Spectrom.* **1999**, *13,* 73−78.
(16) Fenselau, C.; Demirev, P. A. *Mass Spectrom. Rev.* **2001**, *20,* 157−171.
(17) Lay, J. O., Jr. *Mass Spectrom. Rev.* **2001**, *20,* 172−194.
(18) Ramirez, J.; Fenselau, C. *J. Mass Spectrom.* **2001**, *36,* 929−936.
(19) Whiteaker, J.; Karns, J.; Fenselau, C.; Perdue, M. L. **2004**, *1*, 185−194.
(20) Mandrell, R. E.; Harden, L. A.; Bates, A. H.; Miller, W. G.; Haddon, W. F.; Fagerquist, C. K. *Appl. Environ. Microbiol.* **2005**, *71,* 6292−6307.
(21) Fagerquist, C. K.; Miller, W. G.; Harden, L. A.; Bates, A. H.; Vensel, W. H.; Wang, G.; Mandrell, R. E. *Anal. Chem.* **2005**, *77,* 4897−4907.
(22) Fagerquist, C. K.; Bates, A. H.; Heath, S.; King, B. C.; Garbus, B. R.; Harden, L. A.; Miller, W. G. *J. Proteome Res.* **2006**, *5,* 2527−2538.
(23) Jarmon, K. H.; Cebula, S. T.; Saenz, A. J.; Petersen, C. E.; Valentine, N. B.; Kingsley, M. T.; Wahl, K. L. *Anal. Chem.* **2000**, *72,* 1217−1223.
(24) Wahl, K. L.; Wunschel, S. C.; Jarman, K. H.; Valentine, N. B.; Petersen, C. E.; Kingsley, M. T.; Zartolas, K. A.; Saenz, A. J. *Anal. Chem.* **2002**, *74,* 6191−6199.
(25) Demirev, P. A.; Ho, Y.-P.; Ryzhov, V.; Fenselau, C. *Anal. Chem.* **1999**, *71*, 2732−2738. Peneda, F. J.; Lin, J. S.; Fenselau, C.; Demirev, P. A. *Anal. Chem.* **2000**, *72,* 3739−3744.
(26) Demirev, P. A.; Lin, J. S.; Peneda, F. J.; Fenselau, C. *Anal. Chem.* **2001**, *73,* 4566−4573.
(27) Yao, Z.-P.; Demirev, P. A.; Fenselau, C. *Anal. Chem.* **2002**, *74,* 2529−2534.
(28) Peneda, F. J.; Antoine, M. D.; Demirev, P. A.; Feldman, A. B.; Jackman, J.; Longenecker, M. Lin, J. S. *Anal. Chem.* **2003**, *75*, 3817−3822.
(29) Centers for Disease Control and Prevention, Coordinating Center for Infectious Diseases/Division of Bacterial and Mycotic Diseases. http://www.cdc.gov/ncidod/dbmd/diseaseinfo/campylobacter_t.htm (October 6, 2005).
(30) Fouts, D. E.; Mongodin, E. F.; Mandrell, R. E.; Miller, W. G.; Rasko, D. A.; Ravel, J.; Brinkac, L. M.; DeBoy, R. T.; Parker, C. T.; Daugherty, S. C.; Dodson, R. J.; Durkin, A. S.; Madupu, R.; Sullivan, S. A.; Shetty, J. U.; Ayodeji, M. A.; Shvartsbeyn, A.; Schatz, M. C.; Badger, J. H.; Fraser, C. M.; Nelson, K. E. *PloS. Biol.* **2005**, *3*(e15), 72−85.
(31) Miller, W. G.; On, S. L. W.; Wang, G.; Fontanoz, S.; Lastovica, A. J.; Mandrell, R. E. *J. Clin. Microbiol.* **2005**, *43,* 2315−2329.
(32) Boehm, A. M.; Galvin, R. P.; Sickmann, A. *BMC Bioinformatics* **2004**, *5,* 162.
(33) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20,* 3551−3557.
(34) Craig, R.; Beavis, R. C. *Bioinformatics* **2004**, *20,* 1466−1467.
(35) Hirel, P. H.; Schmitter, J. M.; Dessen, P.; Fayet, G.; Blanquet, S. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86,* 8247−8251.
(36) Gonzalez, T.; Baudouy, J. *J. FEMS Microbiol. Rev.* **1996**, *18,* 319−334.
(37) Solbiati, J.; Chapman-Smith, A.; Miller, J.; Miller, Chapter; Cronan, J., Jr. *J. Mol. Biol.* **1999**, *290,* 607−614.

PR060700O